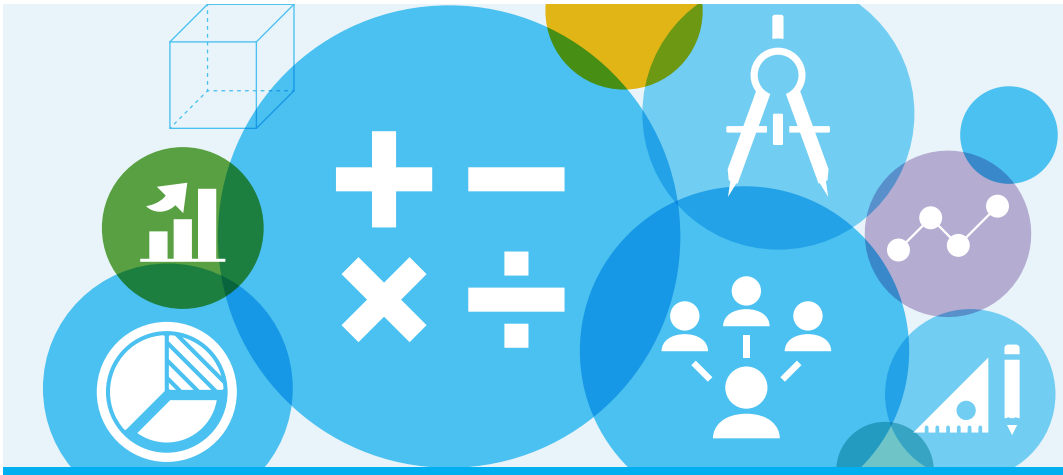


STUDENT VERSION

ANALYZING RELATIONSHIPS: MARRIAGE, DIVORCE, AND LINEAR REGRESSION

Activity Item

The following item is part of this activity and appears at the end of this student version.

- Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia

Student Learning Objectives

- I will be able to assess how well a linear model fits the data by plotting and analyzing the residuals.
- I will be able to determine the impact of outliers on the linear model.
- I will be able to explain the meaning of the slope and of the y-intercept of the linear model in the context of the data.

NAME:_____ DATE:_____

Part 1 – Explore Data in Scatter and Residual Plots

1. Write down the question your class will answer by examining census data.

2. Which data from **Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of the Columbia** will help you answer your question?

3. Make a scatter plot of your data on the grid below.

4. Consider which of your variables you plotted on the horizontal and vertical axes. Did your choice matter in this situation? Why or why not?

5. Choose any data point on your scatter plot to examine. What does that point represent in the context of the data set?

   a. Do the two variables appear to have an association?

   b. If so, is it linear? Weak or strong? Positive or negative? How do you know?

6. Add a line of best fit to your scatter plot and use your graphing calculator to determine the equation for that line. Keep in mind that statisticians typically replace the generic variables x and y with the actual variable names, or at least descriptive variable names. They also place a " ^ " symbol over the y variable to indicate that it is a predicted value. For example, an equation relating 2010 marriage rates for women (as the y variable) and for men (as the x variable) might be: $\widehat{\text{women}} = 0.94 \cdot \text{men} + 0.08$. Write your equation below.

7. A residual is the difference between the actual *y* coordinate of a data point and what the linear model predicts (actual - predicted). Calculate the residual for any two points on your scatter plot, using your line of best fit to determine predicted values. Show your work.

- **Point 1:**

- **Point 2:**

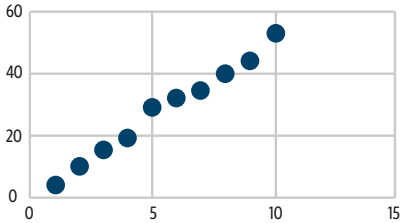
8. If the residual is positive, what does that mean and how does it compare with the line of best fit? What about if the residual is negative?

9. What does it mean if the residual is close to 0? What if the residual isn't close to 0? How would each of these residuals appear in a scatter plot?

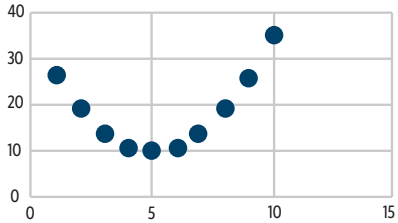
10. You can determine whether a linear model is appropriate for your data by looking at the residual plot, which is a scatter plot of all the residuals versus the *x* variable. The residual plot shows the relationship between the actual data after accounting for the linear pattern. If a linear model is appropriate for the data, there should be no remaining pattern in a residual plot and the points should be spread equally above and below the horizontal line at 0.

    With this information in mind, draw a line to connect each of the three scatter plots with what you think is its residual plot below.

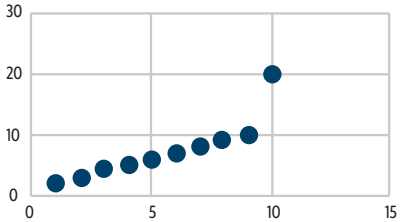
Scatter Plots



A

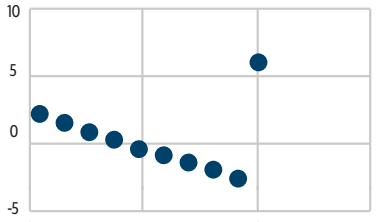


B

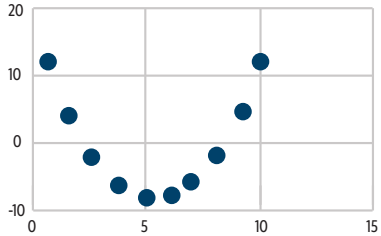


C

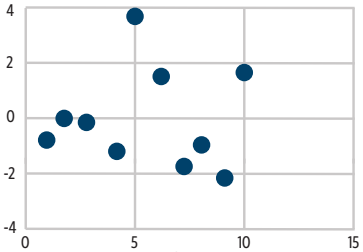
Residual Plots



A



B



C

11. Using the data from **Item 1** and your line of best fit from question 6, make a residual plot in the grid below. Draw a horizontal line at 0, and graph the residual for each data point above or below this line, depending on whether it is positive or negative.

12. Is a linear model appropriate for your data? Explain.

13. You already know that the slope of a linear model represents the rate of change of the line.

    Statisticians talk about slope in slightly different terms: as a measure of comparison between two *x* values. In the 2010 example equation from question 6 ($\widehat{women} = 0.94 \cdot men + 0.08$), the slope is 0.94. This means that if state A has 10 more marriages per 1,000 men than state B, then state A is likely to have approximately 9.4 more marriages per 1,000 women (i.e., 10*0.94) than state B.

    What is the slope of your line of best fit? Explain what the slope means in the context of comparing two points in your data set.

14. You also already know that the y-intercept represents the predicted *y* value when the *x* value equals 0. Use that knowledge to answer the following questions and prompts.
    a. What is the y-intercept for your line of best fit, rounded to the nearest hundredth?

    b. Explain what the y-intercept means in the context of your data.

    c. Does it make sense to interpret the y-intercept in this way? Explain.

15. Some scatter plots contain outliers, which are data points that fall outside of the pattern of the rest of the points (i.e., far from the line of best fit). Looking at both the scatter plot and the residual plot you created, can you identify any outliers? Be sure to explain how you used your residual plot to confirm any outliers.

16. If your scatter plot contains an outlier, identify which U.S. state (or the District of Columbia) that point represents. (If your scatter plot has multiple outliers, pick the greatest one.) Why do you think this location displays this unusual behavior in the context of the data? If your scatter plot does not contain an outlier, explain how you know that.

17. Sometimes an outlier can influence the linear model by pulling the line of best fit toward it. In some cases, outliers make the association appear stronger; in other cases, it appears weaker. One way to determine whether an outlier is influencing the linear model is to remove it from the data set and see what happens.

    If your scatter plot has any outliers, remove these data points from the set and recalculate the new line of best fit. Write it below. Does the removal of outliers change the line significantly? Why or why not?

18. How do the components of your new line of best fit and its equation compare with those of the original line and equation?

19. Put the point you removed in question 17 back in the data set and select another point, close to the middle of your data set, to remove. Recalculate the equation for this line of best fit and write it below. How does this new line compare with the original? How does it compare with the line you calculated for question 17, and why do you think that is?

Part 2 – Make Predictions

Linear models can be used to make predictions. With that in mind, predict the 2010 marriage rate for women in Puerto Rico using the example equation from earlier ($\widehat{women} = 0.94 \cdot men + 0.08$) and the following information: The 2010 marriage rate for men in Puerto Rico is 10.7 marriages per 1,000 men.

1. Write the new equation that you will use to make your prediction. What value does the linear model predict?

2. The actual 2010 marriage rate for women in Puerto Rico is 9.2 marriages per 1,000 women. How close was your prediction in question 1 to the actual value? And how confident are you in this prediction?

Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia

State	Marriage Rates				Divorce Rates			
	2014		2008		2014		2008	
	Women	Men	Women	Men	Women	Men	Women	Men
Alabama	17.0	19.4	20.7	22.5	10.4	9.4	13.5	13.7
Alaska	28.7	29.4	24.8	27.8	10.3	9.1	12.5	13.2
Arizona	17.3	17.7	16.8	18.6	8.8	7.5	12.0	11.3
Arkansas	20.5	19.7	24.5	25.6	11.6	12.0	14.1	11.1
California	17.1	18.3	17.8	19.5	7.2	6.7	9.9	8.3
Colorado	21.0	22.1	20.9	21.9	10.5	9.8	11.7	13.0
Connecticut	13.5	14.1	15.6	17.4	8.0	7.3	8.5	9.3
Delaware	19.4	17.7	14.0	15.8	8.8	10.3	11.8	11.1
District of Columbia	23.6	30.4	17.3	19.5	8.6	5.8	8.4	12.1
Florida	13.7	15.4	17.2	18.1	9.5	7.6	10.6	10.1
Georgia	18.5	19.5	18.9	20.5	11.2	8.7	11.7	11.3
Hawaii	18.4	20.9	17.6	19.2	5.9	6.9	7.7	8.4
Idaho	22.3	22.7	25.5	26.6	14.4	7.4	14.5	10.3
Illinois	15.1	16.5	15.1	17.5	7.3	6.9	9.7	8.7
Indiana	19.3	19.9	18.4	20.4	11.3	10.1	11.8	12.4
Iowa	17.6	18.9	19.6	19.4	6.6	7.5	9.8	9.5
Kansas	19.4	20.7	22.0	24.7	10.8	9.9	12.8	12.6
Kentucky	17.3	18.2	19.9	22.4	11.0	10.6	15.0	13.3
Louisiana	18.0	18.4	19.4	21.9	6.8	8.3	10.4	10.4
Maine	16.4	15.0	16.4	16.9	11.0	10.7	11.9	12.5

Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia (Continued)

State	Marriage Rates				Divorce Rates			
	2014		2008		2014		2008	
	Women	Men	Women	Men	Women	Men	Women	Men
Maryland	14.7	16.0	18.2	20.4	8.2	6.8	10.0	10.1
Massachusetts	14.3	14.6	13.4	15.1	7.0	6.5	8.8	7.9
Michigan	14.1	14.9	15.7	16.7	9.2	8.5	9.9	10.3
Minnesota	16.9	17.9	18.0	18.7	7.8	7.6	9.3	9.3
Mississippi	19.2	20.7	19.1	21.6	10.0	11.3	11.6	14.0
Missouri	16.7	17.2	18.8	21.2	8.4	9.7	11.3	11.2
Montana	16.6	16.5	18.4	18.7	6.9	12.2	10.3	9.6
Nebraska	16.7	18.0	23.8	25.3	8.9	8.6	9.7	12.3
Nevada	21.2	21.3	21.9	21.6	10.4	9.8	11.2	12.8
New Hampshire	18.6	18.5	15.7	17.5	8.8	9.1	9.2	9.8
New Jersey	13.7	15.9	15.4	17.5	5.5	6.6	8.7	7.0
New Mexico	16.9	17.8	21.7	23.0	9.7	8.5	10.2	9.6
New York	14.4	16.7	14.7	17.1	6.3	5.9	7.1	7.0
North Carolina	17.1	18.8	18.5	21.1	9.0	7.8	10.1	9.5
North Dakota	18.5	18.8	23.9	26.3	7.6	5.0	6.0	10.2
Ohio	15.2	16.1	16.5	18.0	8.7	7.8	9.9	9.8
Oklahoma	22.6	23.5	23.0	24.4	12.5	11.6	16.5	13.6
Oregon	18.3	18.4	19.3	20.5	10.4	8.7	11.3	10.4
Pennsylvania	15.0	16.2	14.3	16.0	7.2	7.2	8.0	8.2
Rhode Island	13.9	16.6	13.9	16.5	8.6	6.5	8.5	8.4

Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia (Continued)

State	Marriage Rates				Divorce Rates			
	2014		2008		2014		2008	
	Women	Men	Women	Men	Women	Men	Women	Men
South Carolina	16.8	19.5	15.4	15.7	8.9	8.9	9.8	9.5
South Dakota	21.1	22.7	17.2	21.9	7.6	7.1	8.9	11.2
Tennessee	18.6	19.8	18.2	20.3	9.5	11.0	12.8	12.8
Texas	19.5	21.1	20.6	22.2	10.2	9.8	12.7	10.9
Utah	26.4	24.3	28.4	27.6	11.9	9.5	14.0	12.1
Vermont	14.9	15.2	14.7	15.8	6.7	6.7	10.1	11.6
Virginia	18.9	19.6	18.5	21.0	8.7	8.1	11.2	9.7
Washington	19.8	19.6	21.0	22.5	8.7	8.8	12.5	11.5
West Virginia	17.0	18.1	17.4	19.6	12.4	9.5	12.8	10.3
Wisconsin	16.4	16.1	18.4	18.8	7.9	7.5	9.8	8.5
Wyoming	26.4	24.6	28.1	27.5	8.1	6.8	10.3	21.0

Note: These are survey data and do not necessarily reflect marriages and divorces that took place in a particular state. Instead, they reflect whether people living in a particular state at the time of the American Community Survey interview had divorced or married within the 12 months prior to the interview.